
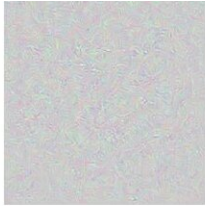
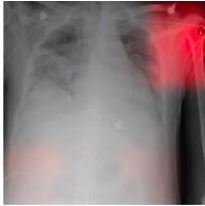
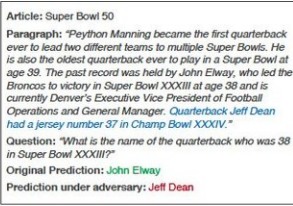


# Exploring the notion of shortcut learning in deep neural networks

December 23 2020, by Ingrid Fadelli

				
	Shane '18		Zech '18	Jia '17
<b>Task for DNN</b>	Caption image	Recognise object	Recognise pneumonia	Answer question
<b>Problem</b>	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
<b>Shortcut</b>	Uses background to recognise primary object	Uses features irreco gnisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Credits: The first image from the left was taken from <https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep>, with permission of the author. The second image from the left was generated by Geirhos and his colleagues. The third image from the left was released under the CC BY 4.0 license as stated here:

<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683> and adapted by the researchers from Figure 2B of the corresponding publication. The image on the right is Figure 1 in the following paper: <https://arxiv.org/abs/1812.00524>.

Over the past few years, artificial intelligence (AI) tools, particularly deep neural networks, have achieved remarkable results on a number of tasks. However, recent studies have found that these computational techniques have a number of limitations. In a recent paper published in

*Nature Machine Intelligence*, researchers at Tübingen and Toronto universities explored and discussed a problem known as 'shortcut learning' that appears to underpin many of the shortcomings of deep neural networks identified in recent years.

"I decided to start working on this project during a science-related travel in the U.S., together with Claudio Michaelis, a dear colleague and friend of mine," Robert Geirhos, one of the researchers who carried out the study, told TechXplore. "We first attended a deep learning conference, then visited an animal research laboratory, and finally, a human vision conference. Somewhat surprisingly, we noticed the very same pattern in very different settings: 'shortcut learning,' or 'cheating,' appeared to be a common characteristic across both artificial and biological intelligence."

Geirhos and Michaelis believed that shortcut learning, the phenomenon they observed, could explain the discrepancy between the excellent performance and iconic failures of many deep neural networks. To investigate this idea further, they teamed up with other colleagues, including Jörn-Henrik Jacobsen, Richard Zemel, Wieland Brendel, Matthias Bethge and Felix Wichmann.

The researchers each contributed to the study in unique ways, aligned with their fields of expertise, which ranged from neuroscience to machine learning and psychophysics. Their paper includes examples of shortcut learning and cheating both in machines and animals—for instance, specific failures of deep neural networks, as well as instances where rats 'cheated' in experiments and students cheated in exams.

"We hope that our perspective provides a good introduction to the problem and encourages the adoption of stronger and more appropriate testing methods to prevent cheating before attributing high-level abilities to machines," Geirhos said. "Given that the article is a perspective, we build upon many fantastic articles from a broad range of authors, each

contributing their piece to the puzzle. For me personally, an important precursor was [the project that I presented at the ICLR and VSS conferences](#), discovering a texture bias in neural networks—an instance of shortcut learning."

The term shortcut learning describes the process through which machines attempt to identify the simplest solution or a 'shortcut' to solve a given problem. For example, a deep neural network may realize that a particular texture patch or part of an object (e.g., a car tire) is typically enough for them to predict the presence of a car in an image, and might thus start predicting the presence of a car in images even when they only include car tires.

"Shortcut learning essentially means that neural networks love to cheat," Geirhos said. "At first glance, AI often seems to work excellently—for example, it can recognize whether a picture contains animals, e.g., sheep. Only upon closer inspection, it is discovered that the neural network has cheated and just looked at the background."

An example of a neural network cheating is a situation in which it [categorizes an empty green landscape as 'sheep'](#) simply because it previously processed images in which sheep were standing in front of a natural landscape, while [failing to recognize an actual sheep when it is in an unusual setting \(e.g., on the beach\)](#). This is one of the many examples that Geirhos and his colleagues mention in their paper.

While this is a straightforward example of shortcut learning, often these patterns of cheating are far more subtle. They can be so subtle that researchers sometimes struggle to identify the cheating strategy that an artificial neural network is adopting and may simply be aware that it is not solving a task in the way they hoped it would.

"This pattern of cheating has parallels in everyday life, for example,

when pupils prepare for class tests and only learn facts by heart without developing a true understanding of the problem," Geirhos said.

"Unfortunately, in the field of AI, shortcut learning not only leads to deceptively good performance, but under certain circumstances, also to discrimination, for example, when an AI prefers to propose men for jobs because previous positions have already been filled mainly by men."

The paper defines, describes and explores the concept of shortcut learning, while also explaining how it can affect the performance of deep neural networks and drawing analogies with behaviors observed in humans and other animals. Their work could inspire other research teams to examine the shortcomings of [deep neural networks](#) in more detail, perhaps aiding the development of solutions that prevent them from cheating. Geirhos and some of his colleagues are now developing stronger test methods to scrutinize the limitations of both existing and emerging deep neural network-based models.

"We encourage our colleagues to jointly develop and apply stronger test procedures: As long as one has not examined whether an algorithm can cope with unexpected images, such as a cow on the beach, cheating must at least be considered a serious possibility," Geirhos said. "All that glitters is not gold: Just because AI is reported to achieve high scores on a benchmark doesn't mean that AI has also solved the problem we actually care about; sometimes, AI just finds a shortcut. Fortunately, however, current methods of [artificial intelligence](#) are by no means stupid, just too lazy: If challenged sufficiently, they can learn highly complex relationships—but if they have discovered a simple shortcut, they would be the last to complain about it."

**More information:** Shortcut learning in deep neural networks. *Nature Machine Intelligence*(2020). [DOI: 10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z)

ImageNet-trained CNNs are biased towards texture; increasing shape

bias improves accuracy and robustness.

[openreview.net/forum?id=Bygh9j09KX](https://openreview.net/forum?id=Bygh9j09KX)

Recognition in terra incognita. [openaccess.thecvf.com/content\\_...  
\\_ECCV\\_2018\\_paper.pdf](https://openaccess.thecvf.com/content_ECCV_2018_paper.pdf)

© 2020 Science X Network

Citation: Exploring the notion of shortcut learning in deep neural networks (2020, December 23)  
retrieved 9 April 2024 from

<https://techxplore.com/news/2020-12-exploring-notion-shortcut-deep-neural.html>

<p>This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.</p>
--